

METHODS AND SYSTEM FOR MULTI-DRUG TREATMENT DISCOVERY

FIELD OF THE INVENTION

[0001] The present invention relates to the field of disease treatment screening and disease treatment discovery, and more particularly to methods and systems for prospectively screening and finding multiple treatments that may be used more effectively than a single treatment approach to treating disease.

BACKGROUND OF THE INVENTION

[0002] The current state of drug discovery techniques largely focus on searching for a single drug which will effectively treat a disease. In the past, such techniques have been successful for diseases which are “locked in” to a single pathway for survival and/or regeneration within its host. For example, penicillin was discovered to be successful in treating many forms of bacterial diseases. However, for more complex diseases, such as the various cancers; viral diseases, such as HIV (AIDS), SARS, and others; and drug-resistant bacteria, such as tuberculosis and others, single drug treatments have not proven effective. Even for drugs which are approved as effective against such diseases, on average, such drugs will only be effective on about thirty percent of patients to which the drug is administered. Indeed, over a prolonged time scale, even penicillin-resistant bacteria have begun to emerge.

[0003] One way to explain this limited efficacy is that complex diseases may have alternate pathways along which they survive and regenerate. Those diseases which get locked into a single pathway of survival by single nucleotide polymorphism (SNP) profile can be effectively treated by a single drug that blocks the single pathway. However, some diseases, like cancer and HIV, for example, are flexible enough, so that if their primary path, as determined by the host SNP profile, is blocked using a single drug, after some time, the disease is able to change its survival path from the primary pathway to an alternate survival pathway. A multi-drug approach to these types of diseases is needed in order to block multiple survival pathways.

- [0004] Although there have been some instances of multi-drug treatments of disease, such instances have been limited and disorganized, in a hindsight fashion, where there is a history of some effectiveness of two or more drugs, when each is used individually to treat a disease. The approach has then been to try the drugs together to see if an improved result can be achieved. One such example of this is the treatment of HIV by a “cocktail” approach. Another example of a multi-treatment approach, which has also been implemented only through a hindsight trial approach after experiencing some success with each treatment individually, is treatment of certain types of cancer with both radiation treatment and drug treatment. In addition to the hindsight track toward developing this regimen, it is noted that such treatments are also generally administered sequentially, in a time-staggered fashion.
- [0005] There is a need to develop organized, forward-looking ways of identifying treatment combinations for potential use together in the treatment of a disease. There is a need to treat diseases, such as viral diseases, in a complex way, because the organisms causing the diseases are very complex, able to mutate and/or use other mechanisms to survive along a different pathway when one pathway is cut off by a single treatment. Likewise, cancers, and some of the other more complex diseases that have been studied for a long time, with no cure found to date, may find more successful treatment regimens when treated multidimensionally. One-dimensional approaches (i.e., as addressed by a single type of treatment) to treatment of many of the more complex diseases have been unsuccessful to date, but generally the big pharmaceutical companies currently continue in their quests to find a “silver bullet”, i.e., a single drug, to cure a disease.
- [0006] What is needed are forward looking screening and discovery techniques for identifying treatments that may be used in combination for treatment of disease. Hence, treatments in combination can create the requisite complexity to challenge sophisticated diseases. There needs to be a strategy, for using foresight in developing multi-treatment approaches, and to move away from the paradigm of drug treatment, radiation treatment or other types of treatments in one-dimensional space, by thinking in multi-dimensional space, to provide treatments that act multi-dimensionally.

[0007] As noted, multiple treatments that are currently used are results of hindsight combinations. For example, it was known that the treatment of liver disease with interferon was shown to be effective, and that ribovirin could also show good results. By combining these treatments, after knowledge of their individual results, it was found that using ribovirin with interferon was more effective. Typically, drug interactions have been approached as a need to identify them to avoid them. What is needed is to look at drugs, as well as other possible types of treatment, to identify them as an advantage against disease, i.e., to identify positive interactions among multiple treatments.

SUMMARY OF THE INVENTION

[0008] The present invention provides methods, systems and recordable media for using expression data characterizing protein pathways of diseases to link relationships between treatment responses of diseased tissues to treatments applied thereto, and expression profiles of the diseased tissues as measured when untreated.

[0009] Methods, systems and recordable media are provided for screening a combination of treatments to specifically target a disease process. Differential expression levels of diseased tissues relative to at least one reference tissue are provided, either by data from previous processing of such tissues, or the tissues may be processed as part of the current procedure, using microarray technology to obtain the differential expression levels. Phenotypic/genotypic differential expression signatures or profiles are provided (either supplied as data which has already been generated, or calculated/generated from the differential expression data provided from an existing data source, or generated at the time of processing) for respective features of respective microarrays for each diseased tissue sample, relative to the at least one reference level. This gives the capability of generating a phenotypic/genotypic signature for each feature on a microarray, across the total number of diseased-tissue samples being studied.

[0010] The diseased tissues are treated with a treatment, and a treatment-response value with respect to each of the diseased tissue samples as effected by the treatment is recorded. The treatment-response vectors are used to generate a vector of phenotypic signatures representing the treatment-response vector of

each of the diseased tissue samples to that treatment which was applied. Any number of treatments or treatment combinations can be successively and independently applied to treat the diseased tissues, so as to generate a vector of phenotypic signatures representing the treatment-response vector of each of the diseased tissue samples, for each treatment or combination of treatments applied.

[0011] The phenotypic/genotypic signatures of the differential expression levels and the phenotypic signatures of the treatment-response values are considered together after appropriate normalization to perform a clustering operation. Upon identifying clusters of related normalized signatures, compounds are characterized by the treatment-response phenotypic signatures caused by those treatments, and which are clustered with phenotypic/genotypic signatures representing differential expression levels representative of the diseased tissue samples. By selecting treatments linked to treatment-response signatures that are associated with different groups of genes, as identified by differential expression signatures, a multi-treatment set of treatments (which set may also include one or more individual members made up of multiple treatments, for example) can be chosen to specifically target genes involved in the disease process over a wide spectrum of protein activity. Such an approach should effectively shut off multiple pathways of survival of the disease process with minimal side effects.

[0012] Further, the present invention may be used to augment an original or existing treatment or treatment combination with a treatment or treatments that cover gene activity of a disease not addressed by the original/existing treatment or treatment combination. Still further, microarray analyses of diseased tissue samples, when dosed by the original/existing treatment may be performed to better expose disease-gene activity for which additional new treatments will be needed to impact those areas not currently impacted by the original/existing treatment.

[0013] These and other advantages and features of the invention will become apparent to those persons skilled in the art upon reading the details of the present invention as more fully described below.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0014] Fig. 1 shows a matrix of tissue sample responses to treatment by various treatments.
- [0015] Fig. 2 show a correlation matrix generated using the treatment profiles in response to the various treatments as shown in Fig. 1 and mRNA expression profiles from microarray experiments on the diseased tissue samples of Fig. 1. Optionally, the set of tissue samples may or may not be biologically conditioned and/or under treatment according to a specified protocol and/or treatment regimen.
- [0016] Fig. 3 schematically illustrates a “black box” model which relates the genotypic signature of a treatment response to the phenotypic signature via a transfer function. More generally, the black box may represent a multiple-input/multiple-output (MIMO) transfer function involving groups of genes and multiple reactions to treatment(s) dosage.
- [0017] Fig. 4 shows a matrix which includes genotypic/phenotypic gene signatures along with phenotypic signatures in response to treatments administered to tissue samples.
- [0018] Figs. 5A-5D show an example of a portion of a major cluster that was identified by the present method with various treatments performed on lung cancer tissues.
- [0019] Fig. 6 is a schematic representation of an ellipsoid which represents the plot of a cluster of vectors from a matrix such as the matrix shown in Fig. 4, or the matrix from which the data generated in Fig. 5 was based on. As noted, this is a schematic representation, as, in reality, data ellipsoids are hyperdimensional with complicated radial and angular geometries.
- [0020] Fig. 7 is a schematic flowchart exemplifying steps that may be performed in carrying out a multi-treatment screening method according to one embodiment of the present invention.
- [0021] Fig. 8 is a block diagram illustrating an example of a generic computer system which may be used in implementing the present invention.

DETAILED DESCRIPTION OF THE INVENTION

- [0022] Before the present methods and systems are described, it is to be understood that this invention is not limited to particular treatments, drugs, diseases, methods, method steps, statistical methods, hardware or software described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only by the appended claims.
- [0023] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are now described. All publications mentioned herein are incorporated herein by reference to disclose and describe the methods and/or materials in connection with which the publications are cited.
- [0024] It must be noted that as used herein and in the appended claims, the singular forms "a", "and", and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a sample" includes a plurality of such samples and reference to "the microarray" includes reference to one or more microarrays and equivalents thereof known to those skilled in the art, and so forth.
- [0025] The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

DEFINITIONS

- [0026] A "genotype" refers to the actual makeup of one or more genes (DNA) in living tissue. A genotypic signature is a textual or electronic representation that directly identifies the genotype.

- [0027] A “phenotype” is related to a genotype, in that it is some sort of physical expression resulting from a blueprint provided by the genotype. A phenotypic signature is a textual or electronic representation of values representing the expression that defines the phenotype. mRNA expression might be considered to be either a genotype or phenotype, as it is in a gray area where the genotype executes the phenotype. It is referred to herein as genotype/phenotype.
- [0028] A “treatment” refers to the administration of an agent to living tissue (generally a diseased tissue) that has some measurable effect on protein production by that tissue, which effect can be inferred by measurement of gene expression levels of the tissue, using microarray technology. “Treatments” may refer to, but are not limited to drugs, compounds, genetic sequences used to target specific locations of the genetic makeup of the tissue, radiation, heat, cryogenics, or any other kind of application that produces an effect as described above.
- [0029] A “biopolymer” is a polymer of one or more types of repeating units. Biopolymers are typically found in biological systems and particularly include polysaccharides (such as carbohydrates), and peptides (which term is used to include polypeptides and proteins) and polynucleotides as well as their analogs such as those compounds composed of or containing amino acid analogs or non-amino acid groups, or nucleotide analogs or non-nucleotide groups. This includes polynucleotides in which the conventional backbone has been replaced with a non-naturally occurring or synthetic backbone, and nucleic acids (or synthetic or naturally occurring analogs) in which one or more of the conventional bases has been replaced with a group (natural or synthetic) capable of participating in Watson-Crick type hydrogen bonding interactions. Polynucleotides include single or multiple stranded configurations, where one or more of the strands may or may not be completely aligned with another.
- [0030] A “nucleotide” refers to a sub-unit of a nucleic acid and has a phosphate group, a 5 carbon sugar and a nitrogen containing base, as well as functional analogs (whether synthetic or naturally occurring) of such sub-units which in the polymer form (as a polynucleotide) can hybridize with naturally occurring polynucleotides in a sequence specific manner analogous to that of two naturally occurring polynucleotides.. For example, a “biopolymer” includes DNA

(including cDNA), RNA, oligonucleotides, and PNA and other polynucleotides as described in US 5,948,902 and references cited therein (all of which are incorporated herein by reference), regardless of the source. An “oligonucleotide” generally refers to a nucleotide multimer of about 10 to 100 nucleotides in length, while a “polynucleotide” includes a nucleotide multimer having any number of nucleotides. A “biomonomer” references a single unit, which can be linked with the same or other biomonomers to form a biopolymer (for example, a single amino acid or nucleotide with two linking groups one or both of which may have removable protecting groups).

[0031] An “array” or “microarray”, unless a contrary intention appears, includes any one-, two- or three-dimensional arrangement of addressable regions bearing a particular chemical moiety or moieties (for example, biopolymers such as polynucleotide sequences) associated with that region. An array is “addressable” in that it has multiple regions of different moieties (for example, different polynucleotide sequences) such that a region (a “feature” or “spot” of the array) at a particular predetermined location (an “address”) on the array will detect a particular target or class of targets (although a feature may incidentally detect non-targets of that feature). Array features are typically, but need not be, separated by intervening spaces. In the case of an array, the “target” will be referenced as a moiety in a mobile phase (typically fluid), to be detected by probes (“target probes”) which are bound to the substrate at the various regions. However, either of the “target” or “target probes” may be the one which is to be evaluated by the other (thus, either one could be an unknown mixture of polynucleotides to be evaluated by binding with the other). An “array layout” refers to one or more characteristics of the features, such as feature positioning on the substrate, one or more feature dimensions, and an indication of a moiety at a given location. “Hybridizing” and “binding”, with respect to polynucleotides, are used interchangeably. A “pulse jet” is a device which can dispense drops in the formation of an array. Pulse jets operate by delivering a pulse of pressure to liquid adjacent an outlet or orifice such that a drop will be dispensed therefrom (for example, by a piezoelectric or thermoelectric element positioned in a same chamber as the orifice). An array may be blocked into

subarrays which may be hybridized as separate units or hybridized together as one array.

[0032] Any given substrate may carry one, two, four or more or more arrays disposed on a front surface of the substrate. Depending upon the use, any or all of the arrays may be the same or different from one another and each may contain multiple spots or features. A typical array may contain more than ten, more than one hundred, more than one thousand more ten thousand features, or even more than one hundred thousand features, in an area of less than 20 cm² or even less than 10 cm². For example, features may have widths (that is, diameter, for a round spot) in the range from a 10 µm to 1.0 cm. In other embodiments each feature may have a width in the range of 1.0 µm to 1.0 mm, usually 5.0 µm to 500 µm, and more usually 10 µm to 200 µm. Non-round features may have area ranges equivalent to that of circular features with the foregoing width (diameter) ranges. At least some, or all, of the features are of different compositions (for example, when any repeats of each feature composition are excluded the remaining features may account for at least 5%, 10%, or 20% of the total number of features), each feature typically being of a homogeneous composition within the feature. Interfeature areas will typically (but not essentially) be present which do not carry any polynucleotide (or other biopolymer or chemical moiety of a type of which the features are composed). Such interfeature areas typically will be present where the arrays are formed by processes involving drop deposition of reagents but may not be present when, for example, photolithographic array fabrication processes are used,. It will be appreciated though, that the interfeature areas, when present, could be of various sizes and configurations.

[0033] Each array may cover an area of less than 100 cm², or even less than 50 cm², 10 cm² or 1 cm². In many embodiments, the substrate carrying the one or more arrays will be shaped generally as a rectangular solid (although other shapes are possible), having a length of more than 4 mm and less than 1 m, usually more than 4 mm and less than 600 mm, more usually less than 400 mm; a width of more than 4 mm and less than 1 m, usually less than 500 mm and more usually less than 400 mm; and a thickness of more than 0.01 mm and less than 5.0 mm, usually more than 0.1 mm and less than 2 mm and more usually

more than 0.2 and less than 1 mm. With arrays that are read by detecting fluorescence, the substrate may be of a material that emits low fluorescence upon illumination with the excitation light. Additionally in this situation, the substrate may be relatively transparent to reduce the absorption of the incident illuminating laser light and subsequent heating if the focused laser beam travels too slowly over a region. For example, substrate 10 may transmit at least 20%, or 50% (or even at least 70%, 90%, or 95%), of the illuminating light incident on the front as may be measured across the entire integrated spectrum of such illuminating light or alternatively at 532 nm or 633 nm.

- [0034] Arrays can be fabricated using drop deposition from pulse jets of either polynucleotide precursor units (such as monomers) in the case of *in situ* fabrication, or the previously obtained polynucleotide. Such methods are described in detail in, for example, the previously cited references including US 6,242,266, US 6,232,072, US 6,180,351, US 6,171,797, US 6,323,043, U.S. Patent Application Serial No. 09/302,898 filed April 30, 1999 by Caren et al., and the references cited therein. As already mentioned, these references are incorporated herein by reference. Other drop deposition methods can be used for fabrication, as previously described herein. Also, instead of drop deposition methods, photolithographic array fabrication methods may be used. Interfeature areas need not be present particularly when the arrays are made by photolithographic methods as described in those patents.
- [0035] Following receipt by a user, an array will typically be exposed to a sample (for example, a fluorescently labeled polynucleotide or protein containing sample) and the array then read. Reading of the array may be accomplished by illuminating the array and reading the location and intensity of resulting fluorescence at multiple regions on each feature of the array,. For example, a scanner may be used for this purpose which is similar to the AGILENT MICROARRAY SCANNER manufactured by Agilent Technologies, Palo Alto, CA. Other suitable apparatus and methods are described in U.S. patent applications: Serial No. 10/087447 "Reading Dry Chemical Arrays Through The Substrate" by Corson et al.; and in U.S. Patents 6,518,556; 6,486,457; 6,406,849; 6,371,370; 6,355,921; 6,320,196; 6,251,685; and 6,222,664. However, arrays may be read by any other method or apparatus than the

foregoing, with other reading methods including other optical techniques (for example, detecting chemiluminescent or electroluminescent labels) or electrical techniques (where each feature is provided with an electrode to detect hybridization at that feature in a manner disclosed in US 6,251,685, US 6,221,583 and elsewhere). A result obtained from the reading may be used in accordance with the techniques of the present invention in screening and finding multiple drug treatment therapies. A result of the reading (whether further processed or not) may be forwarded (such as by communication) to a remote location if desired, and received there for further use (such as further processing).

- [0036] When one item is indicated as being "remote" from another, this is referenced that the two items are at least in different buildings, and may be at least one mile, ten miles, or at least one hundred miles apart.
- [0037] "Communicating" information references transmitting the data representing that information as electrical signals over a suitable communication channel (for example, a private or public network). "Forwarding" an item refers to any means of getting that item from one location to the next, whether by physically transporting that item or otherwise (where that is possible) and includes, at least in the case of data, physically transporting a medium carrying the data or communicating the data.
- [0038] A "processor" references any hardware and/or software combination which will perform the functions required of it. For example, any processor herein may be a programmable digital microprocessor such as available in the form of a mainframe, server, or personal computer (desktop or portable). Where the processor is programmable, suitable programming can be communicated from a remote location to the processor, or previously saved in a computer program product (such as a portable or fixed computer readable storage medium, whether magnetic, optical or solid state device based). For example, a magnetic or optical disk may carry the programming, and can be read by a suitable disk reader communicating with each processor at its corresponding station.
- [0039] Reference to a singular item, includes the possibility that there are plural of the same items present.
- [0040] "May" means optionally.

[0041] Methods recited herein may be carried out in any order of the recited events which is logically possible, as well as the recited order of events.

[0042] Fig. 1 shows a matrix 100 wherein each column represents one of sixty samples (i.e., cancer cell lines) from the National Cancer Institute that were tested for responses to various drugs, with each of the rows representing a drug that was used to treat the tissues. There are currently about 70,000 drugs to be tested individually against the sixty samples, but not all are shown in the figures, for the purpose of simplifying the explanation. Optionally, one could test specific combinations of compounds on the cell lines. Further, as noted above, other types of treatments (e.g., radiation at various levels, other compounds, genetic sequences, and/or the like) could be applied and tested in addition to the drugs shown, or alternatively to this test. Also in this specific application, only one response to an application of a combination of treatments/compounds is measured. Thus, row 1 shows the phenotypic signature of the response of the sixty tissues to treatment/drug 1. The phenotypic signature includes a treatment (in this case, drug) response value related to expressed levels of proteins used by the disease, e.g., concentration of drug (or level of radiation, or amount or dose of some other treatment) required to inhibit tumor growth by fifty percent over a specified time interval, referred to as GI_{50} , for each of the sixty cells (tissue samples) to which the treatment/drug was applied. Different techniques may be used for the treatment and response measurement. One approach is to determine how much of the treatment needs to be applied to retard cell growth by fifty percent over a fixed amount of time (GI_{50}). Another approach is to apply a fixed amount of the treatment/drug and see how much inhibition results over a fixed amount of time.

[0043] The phenotypic response gives a signature of the variation in the response to the treatment/drug as it impacts disease-active proteins, the levels of which vary across the samples. The mRNA transcription sequences (from the genes, which range up to about 30,000) may number from about 100,000 to 200,000 varieties as produced by gene mutations, varying transcription factors, and/or splice variants and other factors modifying the basic gene message. The phenotypic signatures across the sixty samples are thus induced by level variations of disease-specific proteins as produced by a subset of the 100,000 to 200,000 possible transcription signatures as modified by translation factors specific to each of the sixty cell-line samples. Such response profiles are measured for each of about 70,000 candidate treatments. Each treatment-free mRNA signature is ratioed to its corresponding baseline reference signature to produce a differential-expression profile to characterize disease activity. Thus, after appropriate transformation and normalization, each treatment-induced phenotypic signature in Fig. 1 are compared with differential-expression signatures from genes measured on a single two-color microarray or two single-color microarrays, to map compounds to the transcriptome locations of the diseased tissue samples. Additionally, replicates may be run of each tissue sample against each treatment, to perform an average response with regard to each, to reduce the noise in the phenotypic signatures. Likewise, replicate arrays may be run to reduce noise in gene expression readings. Additionally or alternatively, phenotypic signatures may be processed using “self-self” prediction techniques such as described in co-pending, commonly owned Application Serial No. (not yet assigned, Attorney’s Docket No. 10030281-2) filed March 27, 2003 and titled “Method and System for Predicting Multi-Variable Outcomes”, and in Application Serial No. 60/368,586 filed March 29, 2002 and titled “Generalized Similarity Least Squares Predictor”, both of which are incorporated, in their entireties, by reference thereto.

[0044] Thus, in Fig. 1, each of the tissues is run once treatment-free on a microarray analysis, to create a basis for screening all treatments. For example, there may be 30,000 features on a microarray used to run the analysis. This produces 30,000 signatures, across the sixty sample tissues, to be compared with each overall treatment-induced phenotypic signature shown in Fig. 1. A typical approach to try and find relationships among data on this order of magnitude would be to produce a correlation matrix 200 on the order of about 70,000 x 30,000 values. For example, the 70,000 treatments are identified over the columns of the matrix 200 shown in Fig. 2, with 30,000 mRNA values identified in the rows of the matrix 200. Such a correlation matrix 200 is generated by calculating the inner product of the values of the 70,000 treatment phenotypes (see Fig. 1) with the 30,000 mRNA signatures and typically centered by means and scaled by standard deviation. The inner product may be calculated as follows:

$$\frac{\overrightarrow{D_j} \bullet \overrightarrow{mRNA_k}}{N} = i_{j,k}$$

where

$\overrightarrow{D_j}$ is the vector representing the phenotypic expression across samples when treated with the j^{th} treatment in the list, where j ranges from 1 to 70,000 treatments, in this example;

$\overrightarrow{mRNA_k}$ is the vector representing the sample profile of differential-expression for the k^{th} mRNA sequence (gene), as measured from microarray analysis(es), where k ranges from 1 to 30,000, in this example;

N is the number of samples, e.g., 60 cell lines; and

$i_{j,k}$ is the value in correlation matrix 200 filling the j^{th} column and the k^{th} row.

[0045] Thus, the values in matrix 200 are a cross product of normalized or standardized vectors, and optionally centering by subtracting a vector mean value. Once matrix 200 is produced, then various clustering techniques may be applied to try and identify blocking patterns in the data, or “clusters” which might indicate that certain treatments are effecting certain groups of mRNA. Alternatively, clustering techniques may be applied to the treatment and mRNA matrices prior to performing the cross product operation, in an effort to reduce the sizes to something less than 70,000 and 30,000 vectors, respectively.

[0046] The present inventor believes that important information is dropped or lost when such data is processed into a correlation matrix, such as by performing a cross product/inner product procedure to obtain matrix 200, as described above. When such an operation is performed, phase information is lost, as is well-known, particularly among electrical engineers. Because the phenotypic responses to the treatments at issue are measured as the output (i.e., phenotypic signature) and the inputs are the mRNA values, there is a protein link between the input and output, i.e., the mRNA instructs the generation of proteins in response to the disease. Treatments impact disease proteins to produce the phenotypic output response. Thus, the present inventor believes that it is important to consider the phase relationships between the inputs (mRNA) and outputs (phenotypic responses), as important factors based on the protein links between the inputs and outputs.

[0047] Hence, the approach taken by the present invention is akin to a lumped parameters modeling approach. Referring to Fig. 3, the genotypic signature of mRNA is modeled as the input 310, the phenotypic signature is modeled as the output 320, and a “black box” 330 represents a transfer function that transforms input 310 to output 320. Note that input 310 and output 320 may be single inputs and outputs, e.g., a single gene and single drug impact or single protein production, or multiple inputs and outputs, e.g., groups of genes producing groups of proteins and/or multiple drug impacts. The present invention is adapted to further modifications within the black box as additional information becomes known about the relationships between genes and the production of proteins resultant therefrom. This will enable the input and output to be modeled, for example, like an electrical circuit, with black box 330 containing “resistances”, “inductances”, “capacitances”, “emfs”, etc., that model the best knowledge that is gained with regard to the relationships between the genes and proteins produced therefrom, which will enable the modeling of an accurate phase relationship between the mRNA signatures and the phenotypic signatures. Such complete modeling knowledge is currently not known however, and it may be at least five years, or more, before such knowledge is obtained in sufficient detail.

[0048] In the meantime, this technique assumes that an input signature is either in phase (0°), or out of phase (i.e., 180° out of phase) with the output signature. This assumes that a gene is either directly related to (i.e., involved in the production of) a protein that is expressed (in sync, or in phase) or involved in regulating the production of the protein (trying to counteract the expression of the protein, i.e., perfectly out of sync or out of phase). Therefore an “out of phase” signature is generated in addition to the “in phase” signature which is read from the microarray results. Rather than taking a cross product and forming a correlation matrix, this methodology assembles all of the signatures, after appropriate transformation such as by use of Log transforms, into one large list of profiles, properly normalized, as shown in Fig. 4.

[0049] Matrix 400, for the example that we have been discussing, includes sixty columns, one for each tissue sample. The rows of matrix 400 include the mRNA signatures (which may be thought of as genotypic profiles or phenotypic profiles (genotypic/phenotypic profiles)), which in this example includes 60,000 rows (although it could range up to about 400,000 rows, if functional variants are considered separately) since the signatures are represented as “in phase” and “out of phase”, and thus result in two signatures for each of the 30,000 features on the microarray. The “out of phase” signatures are generated by inverting the “in phase” signatures that are read from the microarray. Matrix 400 further includes the phenotypic signatures, which are the responses to the treatments (in this example, 70,000 rows).

[0050] Rather than clustering within a single zone (i.e., the zone containing the 60,000 genotypic/phenotypic mRNA-expression signatures or the zone containing the 70,000 phenotypic drug-induced signatures), or performing a cross product correlation to combine these two zones (thereby losing valuable information, as discussed above), clustering procedures are performed over the entire matrix 400 as a whole. Each of the signatures are normalized for comparison purposes. One method of normalizing used is by Z-scoring, although other normalization methods may be substituted. Also weighting techniques may be applied so that highly upregulated features do not receive overamplified attention during the clustering process. Further, noisy profiles may be weighted with relatively reduced weight.

[0051] Clustering of the information is performed using any cluster technique that is capable of clustering similar signatures, such as K-means or other known techniques. However, it is preferred to perform the clustering using the tools and techniques described in co-pending, commonly owned Application Serial No. 09/986,746 filed November 9, 2001 and titled "System and Method for Dynamic Data Clustering" which application is incorporated herein, in its entirety, by reference thereto. Similarity may be defined by the Euclidean distance between the normalized vectors in matrix 400. These dynamic clustering techniques are scalable, and parallelizable, so that they can handle large scale problems such as those presented in the context of the present invention. The result of these clustering operations gives clusters which contain mixtures of phenotypic signatures (treatment signatures from the 70,000 row zone) and genotypic/phenotypic signatures (expressed disease gene (mRNA) signatures from the 60,000 row zone). Sometimes the clustering occurs among genes in phase and sometimes with genes out of phase. Some cluster may contain only treatment-induced signatures and some clusters may contain only mRNA-expression signatures. The clusters of interest to the present techniques contain both types of signatures.

[0052] Figs. 5A-5D show an example of a portion of a major cluster that was identified by the present method with various treatments performed on lung cancer tissues. The method employed measures the distance that each profile is from the densest part of a high dimensional ellipsoid characterizing the cluster that the profiles are identified as belonging to. The distances are identified under the "DISTANCE" column 510 in the chart shown in Figs. 5A-5D.

[0053] Fig. 6 is a schematic representation of an ellipsoid 600 which represents the plot of a cluster of vectors from a matrix such as matrix 400 or a similar matrix assembled (such as that for the example of Fig. 5), when the vectors are plotted in high dimensional space. The clustering techniques described above are designed to find and identify such clusters, as well as locate the densest part of each ellipsoid cluster, shown as diagonal or “ridge” 610 in Fig. 6. Using a dynamic data clustering system as disclosed in Application Serial No. 09/986,746, the system uses force functions to converge a mathematical probe to the densest part of a profile cluster. Thus, the system not only identifies the clusters but defines a point of reference for all profiles in the cluster, with respect to each cluster identified. The distance of each profile from the center of the cluster it belongs to can be defined by any viable distance metric. An example of a distance metric used is Euclidean distance. The relative angular positions of the profiles may also be consequential for selecting combinations of effective treatments.

[0054] The distance values from the center are identified as “DISTANCE” for the calculations made with regard to the data in the example shown in Figs. 5A-5D. “Sevinon” is noted as being a member of a group of similar treatments that are the closest to the densest location 610 of the ellipsoid in the example of Fig. 5A, as noted by the subgroup 530. Figs. 5A-5D show a portion of the clustering results for a subset of the sixty cell lines discussed above, where the subset were tissue samples that are specific to lung cancer. Figs. 5A-5D show data from only a portion of the most dominant cluster identified by the process (cluster 23). Column 501 (Idrow) identifies the row index of each profile shown in the data. Columns 502-509 contain values that make up the profiles that were identified in the cluster. Column 511 identifies the cluster to which the displayed profiles were found to belong (in this case, cluster 23). Column 510 indicates “DISTANCE”, which is a distance measurement from the densest location 610 in the cluster. Columns 518 and 519 contain treatment identifiers (in this example, drug identifiers) and gene identifiers by name and by NSC identifier (as provided by NIH).

[0055] It can be observed that the data (arranged in rows which characterize profiles, which can be expressed as vectors, plotted on an ellipsoid, as discussed) is arranged in four subgroups in this example, with the first subgroup 530 being those results which are the closest to the densest location 610 of the ellipsoid, subgroup 540 being the next closest subgroup, subgroup 550 being further away from the densest location 610 than subgroup 540, but closer than subgroup 560, and subgroup 560 being the furthest from the densest location 610, out near the periphery of the ellipsoid 600.

[0056] Although the distances measured do not identify angularity between profiles, or which side of the ridge 611(dominant, principal axis running through the densest location) a particular vector (profile) lies on, the distance value does give an indication of how close to the ridge 611 and densest location 610 that vector lies. For example, the distance 512 for "Sevinon", which is located in subgroup 530, is shown close to ridge 610 in Fig. 6. Another treatment vector for "Gemcitabine", located in subgroup 560, was measured 514 to be further from ridge 611 as shown in Fig. 6. A third treatment vector for "Paclitaxel-Taxol" located in the subgroup 550, is shown at a somewhat intermediate distance 516 relative to distances 512 and 514, and the treatment vector for "Mitoxantrone", located in subgroup 540, is shown at an intermediate distance 518 between the distances for 512 and 514.

[0057] The present approach is to find a combination of treatments, such as those represented in Fig. 6, that show relationships to different genes in the clustered profile, so that each treatment appears to be related to different combinations of genes involved in the disease process. The treatments are then combined and tested together, each in a low dose/amount to observe whether a combinatorial synergistic effect on this disease is achieved by the combination. Serendipitously, the low dose/amount combination reduces the side effects of each individual treatment in the combination. Of course any combination where an adverse reaction occurs among two or more treatments in the combination used would be discarded as being unsuitable for use as a potential treatment combination. Useful combinations will specifically and effectively target the disease process being studied.

[0058] Further, by identifying the treatment vectors as in the example shown above, various combinations of treatments in a treatment family (such as, for example, drugs in the drug family, or related compounds in a compound family) of each identified treatment vector may be tested in the manner described above, as related treatments in a treatment family (e.g., drugs in a drug family) will generally fall within similar relative distances from the densest location of the ellipsoid. For example each of the following family members of the Sevinon family may be substituted for Sevinon , that was originally tested in the selected combination of drugs: Declid; Desenex, solution; a component of Desenex; Renselin, Undec-10-enoic acid; Undecyl-10-enic acid; Undecylenic acid; UNDECEN-10- ACID-1; WLN:QV9U1; 10-Hendecenoic; 10-Henedecenoic acid; 10-Henedecenoic acid; 10-Undecenoic acid (8CI9CI); 10-Undecylenic acid; and 9-Undecylenic acid.

[0059] In this way, the combinations of treatments may be predicted to try as potential combinations for multiple treatments in patients which will address the broadest spectrum of genes related to the production of proteins seen as elevated or inhibited when a tissue is in the disease state. By testing the predicted combinations, useful combinations for treatment in patients will be much easier to identify. The present invention is a forward –looking way of choosing and predicting specific combinations of treatments to test, e.g., using high-throughput (HTP) screening of treatment combinations, and as such, greatly reduces the time to finding successful combinations, which currently have only been discovered accidentally, through hindsight and experiences gained through individual treatments.

- [0060] The treatments identified are targeted to the genes involved in the disease process. Because of this, the chances of significant side effects are reduced. For those combinations found to be effective in the sample tissues, further testing, such as animal testing would be warranted to study any effects the treatments may have on normal tissues within an organism, since the testing with the disease tissue samples only proves that the combination of drugs applied is effective at treating the diseased tissues. For the cancer examples discussed above, testing with the tissue samples would only show that the combination of treatments effectively kills the cancer cells, or not. Animal testing would further show the effects of the treatment combination on the normal tissues in the organism, to see if the animal survives the treatment combination.
- [0061] Protein pathways, implicated by differential gene expression levels when comparing treated tissues to non-treated tissue and among diseased and non-diseased tissues are used to produce phase relations between treatment responses and expression profiles across the tissue samples being tested. An example of such is the sixty cancer cell lines referred to above. Using the techniques described above, all phase-related profiles are normalized and clustered. The number and sizes of the resulting clusters may indicate their relative importance as to effective treatments. For example, cluster 23, a portion of which is shown in Figs. 5A-5D was very large and very dominant with respect to other clusters identified. The structure of each cluster infers treatment-gene associations to guide multi-treatment selections.

[0062] Fig. 7 is a schematic flowchart exemplifying steps that may be performed in carrying out a multi-treatment discovery method according to one example of the present invention. At step 710, diseased tissue samples are procured and identified. These samples are for testing treatment hypotheses against. For example, one method performed used sixty cancer cells lines received from the National Cancer Institute. The samples may be next processed (step 720) using microarray technology, without any treatments being applied to get baseline disease phenotypic signatures of proteins expressed by the diseased tissues. Alternatively, processing may begin with baseline disease phenotypic signatures having been already determined and supplied from another source, such as an outside supplier or research facility, for example. The tissue samples may be human or animal. The diseased tissues are run against a reference, such as healthy tissue(s) of the same type, or a pooled reference of normal tissues, or some other established normal sample that establishes a normal baseline of protein productions, using a two-color, two channel array, or by comparing signals from two single-channel arrays. This comparison process determines the excess amounts or protein and /or inhibited amounts of protein being produced in each of the diseased tissues, relative to normal production. These differential expression readings are what are plotted to produce the diseased tissue baseline phenotypic/genotypic signature across the total number of tissue samples. It is also noted that, alternatively, these differential expression reading may be supplied from an already existing source, such as a database containing results from processing already having been previously accomplished.

[0063] Hence, the comparison of the baseline phenotypic signatures of the diseased tissues with baseline phenotypic signatures of normal or reference tissues gives expression ratios of the proteins, absent any impaction by treatments. The expression ratios indicate protein levels that treatments must impact/compensate in order to be effective against the disease. Thus, for each gene, the phenotypic signature of expression for that gene across all of the diseased tissues is determined by microarray processing. That phenotypic signature is compared to a reference of some sort, such as a phenotypic/genotypic signature of a healthy cell sample, or a pooled reference containing an average phenotypic signature of several healthy cell samples, for example. This comparison gives an idea of how much the diseased genes are upregulated or downregulated, in the diseased tissues, without any treatment, i.e., the ambient activity of the disease, with respect to each gene observed.

[0064] Next, the diseased tissue samples are treated with a treatment (e.g., compound, drug, radiation, genetic sequence, or other type of treatment) at step 730, and responses to the treatment are measured with respect to each tissue sample. A phenotypic signature is generated from the measured responses to the treatment at step 740. Typically, a measurement is taken as to what concentration/amount of the treatment is required to reduce the tissue growth by fifty percent over a fixed period of time, or a measurement of how much the tissue growth slows over a fixed period of time given a fixed amount (concentration) is measured. Note, however, that these are only example measurements, and that any measurement that quantitatively relates the impact on the disease-related proteins with the treatment being administered can be used to generate the phenotypic signatures. In the first example given, the concentration values are recorded with regard to each diseased tissue, and these values make up the phenotypic signature for the treatment having been administered. Tissue growth may be measured by volumetrically, by area of a sample, diameter of a sample, mass of a sample, or other quantitative comparison measurement technique, for example.

[0065] In reducing the tissue growth, it is inferred that the treatment blocks or inhibits expression of some of the proteins expressed by the genes. The amount of treatment that it takes to block or inhibit such expression depends upon the

amount of protein that was expressed in each sample when no compound was applied. Therefore, a variation in treatment response across the diseased-tissue samples results because there is a variation in the amount of protein produced in each sample as a response to the disease. The proteins are produced from the gene expression message (mRNA that was measured). By comparing the treatment-phenotypic signature with the baseline diseased tissue expression ratios, a treatment-phenotypic signature which is similar to a phenotypic signature produced by a particular mRNA infers that the treatment that was administered has an impact on a protein produced as a result of the expression of that gene/mRNA. Also, “out of sync” or “out of phase” diseased/baseline signatures which are similar to a treatment-phenotypic signature may infer an effect of that particular treatment on the particular gene/mRNA characterized by the “out of sync” or “out of phase” signature, as a gene may play some role in the production of the protein through direct or indirect inhibition.

[0066] Steps 730 and 740 are repeated to generate phenotypic signatures with regard to each treatment to be tested against the diseased tissue samples. It is noted that each treatment applied during this process need not be of the same type. For example, a radiation treatment may be the next treatment to be processed, followed by an additional drug treatment. Any possible combination of treatment signatures may be generated for screening as described. Further optionally, “out of sync” or “out of phase” phenotypic signatures for the differential expression levels of the ambient state of the diseased tissue may be generated at step 755, for use in comparison as described above, by inverting the signature values measured in step 720. In either case, all phenotypic signatures are considered together, as described above. At step 760, clustering of the signatures is performed (typically after normalizing the profiles (signatures)).

[0067] Analysis of a cluster of profiles is performed to determine candidate treatments that are clustered with particular mRNA(gene) profiles. By selecting a plurality of treatments at step 770, each of which is shown to be similar to a different gene or group of genes (as shown by its plot in the cluster ellipsoid, as well as similarity measurements with respect to ambient diseased tissue phenotypic profiles), and each of which also belongs to the same cluster, a prediction for a multi-treatment of the disease against which the treatments were

tested may be made. A goal of this type of selection for a predicted multi-treatment regimen, is to treat, by directly targeting as many genes as possible that are linked to the production of proteins involved in the growth of the disease tissue. By using this broad-spectrum approach, while at the same time specifically targeting the genes populating the broad spectrum, significantly smaller doses/amounts of each treatment may show combinatorial efficacy, and all pathways along which the disease can survive may be cut off.

[0068] After a candidate group of treatments has been selected at step 770, the plurality of treatments are tested at step 780, initially to rule out a combination if it clearly lacks efficacy, or is unacceptably toxic, for example. One method of testing the plurality of treatments as a combination treatment for an initial reading of efficacy, is to treat the diseased tissue samples with the selected treatment combination and monitor, as described above, to determine what degree, if any, of inhibition of the disease the treatment combination causes. If the degree of inhibition is less than a predetermined amount, then the selected group of treatments fails this test at step 790. Otherwise, it passes, and the selection of treatments may be continued to use for further, more in-depth studies (step 795), such as animal studies, for example.

[0069] Another type of testing that may be performed at step 780 in addition, or alternative to that previously mentioned, is to determine an ADME (Adsorption, Distribution, Metabolism, and Elimination) profile of the combination of treatments or an ADME-tox (Adsorption, Distribution, Metabolism, Elimination and Toxicity) profile. This can be performed by comparing the expression profiles of tissues treated with the selected treatment combination, with gene-expression profiles characterizing the results of treatment of similar tissues with one or more treatments having known toxic characteristics which are unacceptable for use in treatment. By performing similarity comparisons, if the expression results from treatment with the selected treatment combination are closer than a predetermined threshold to expression results from treatment with any known treatment that has unacceptable toxicity, then the selected group of treatments fails this test at step 790. Otherwise, it passes, and the selection of treatments may be continued to use for further, more in-depth studies (step 795), such as animal studies, for example.

- [0070] When a selected group of treatments fails testing at step 790, processing returns to step 770, where another plurality of treatments is chosen, which is different from any previously selected combination, although selected by the same selection criteria.
- [0071] Hence, efficacy and toxicological-pharmacological properties of individual treatments and selected treatment combinations may be predicted by estimation or measurement of their impact on all genes to produce an expression signature for comparison with gene expression profiles of known efficacy and ADME-tox properties using the herein described methodology.
- [0072] An optional technique may be employed while processing according to the techniques described in Fig. 7, to prospectively eliminate, or predict certain treatment combinations likely to prove ineffective or toxic. This technique involves, in addition to those steps described in Fig. 7, generating phenotypic signatures characterizing sample responses to treatments (each preferably individually applied to the samples, like in step 740) having known, unacceptable toxicity characteristics, known ADME-tox profiles which are undesirable, and/or are known to lack efficacy. These profiles/signatures are added to the matrix containing the treatment-induced profiles being tested, as well as the in-phase and out-of phase signatures characterizing the tissue samples when they are not being treated with a treatment. Then, when clustering is performed at step 760, the known signatures are noted as to their relative positions in the resulting clusters. Treatment-induced signatures located at a distance (such as Euclidean distance, for example) from a signature characterizing an undesirable known treatment that is less than a predetermined distance, are considered to be too similar to the undesirable treatment, and are not selected at step 770, but automatically eliminated from further processing. This decreases the time to process such a large number of treatments in the quest to find acceptable treatment combinations.
- [0073] Embodiments of the present invention as described herein employ various process steps involving data stored in or transferred through computer systems. The manipulations performed in implementing this invention are often referred to in terms such as calculating, normalizing, or solving. Any such terms describing the operation of this invention are machine operations. Useful

machines for performing the operations of embodiments of the present invention include general or special purpose digital computers, analog computational devices/computers or other similar devices. In all cases, there is a distinction between the method of operations in operating a computer and the method of computation itself. Embodiments of the present invention relate to method steps for operating a computer in processing electrical or other physical signals to generate other desired physical signals.

[0074] Embodiments of the present invention also relate to an apparatus for performing these operations. This apparatus may be specially constructed for the required purposes, or it may be a general purpose computer selectively activated or reconfigured by a computer program and/or data structure stored in the computer. The processes presented herein are not inherently related to any particular computer or other apparatus. In particular, various general purpose machines may be used with programs written in accordance with the teachings herein, or it may be more convenient to construct a more specialized apparatus to perform the required method steps.

[0075] Fig. 8 illustrates a typical computer system in accordance with an embodiment of the present invention. The computer system 800 includes any number of processors 802 (also referred to as central processing units, or CPUs) that are coupled to storage devices including primary storage 806 (typically a random access memory, or RAM), primary storage 804 (typically a read only memory, or ROM). As is well known in the art, primary storage 804 acts to transfer data and instructions uni-directionally to the CPU and primary storage 806 is used typically to transfer data and instructions in a bi-directional manner. Both of these primary storage devices may include any suitable computer-readable media such as those described above. A mass storage device 808 is also coupled bi-directionally to CPU 802 and provides additional data storage capacity and may include any of the computer-readable media described above. Mass storage device 808 may be used to store programs, data and the like and is typically a secondary storage medium such as a hard disk that is slower than primary storage. It will be appreciated that the information retained within the mass storage device 808, may, in appropriate cases, be incorporated in standard fashion as part of primary storage 806 as virtual memory. A specific mass

storage device such as a CD-ROM 814 may also pass data uni-directionally to the CPU.

[0076] CPU 802 is also coupled to an interface 810 that includes one or more input/output devices such as such as video monitors, track balls, mice, keyboards, microphones, touch-sensitive displays, transducer card readers, magnetic or paper tape readers, tablets, styluses, voice or handwriting recognizers, or other well-known input devices such as, of course, other computers. Finally, CPU 802 optionally may be coupled to a computer or telecommunications network using a network connection as shown generally at 812. With such a network connection, it is contemplated that the CPU might receive information from the network, or might output information to the network in the course of performing the above-described method steps. The above-described devices and materials will be familiar to those of skill in the computer hardware and software arts.

[0077] The hardware elements described above may implement the instructions of multiple software modules for performing the operations of this invention. For example, instructions for clustering vectors may be stored on mass storage device 808 or 814 and executed on CPU 808 in conjunction with primary memory 806.

[0078] In addition, embodiments of the present invention further relate to computer readable media or computer program products that include program instructions and/or data (including data structures) for performing various computer-implemented operations. The media and program instructions may be those specially designed and constructed for the purposes of the present invention, or they may be of the kind well known and available to those having skill in the computer software arts. Examples of computer-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM, CDRW, DVD-ROM, or DVD-RW disks; magneto-optical media such as floptical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM). Examples of program instructions include both machine code, such as produced by a compiler, and files containing higher level code that may be executed by the

computer using an interpreter.

[0079] While the present invention has been described with reference to the specific embodiments thereof, it should be understood by those skilled in the art that various changes may be made and equivalents may be substituted without departing from the true spirit and scope of the invention. In addition, many modifications may be made to adapt a particular situation, treatment, tissue sample, process, process step or steps, to the objective, spirit and scope of the present invention. All such modifications are intended to be within the scope of the claims appended hereto.